Numerical Optimisation:
Line search methods

Marta M. Betcke

m.betcke@ucl.ac.uk

Kiko Rullan

f.rullan@cs.ucl.ac.uk

Department of Computer Science,
Centre for Medical Image Computing,
Centre for Inverse Problems
University College London

Lecture 2 & 3

Descent direction

Descent direction is a vector $p \in \mathbb{R}^n$ for which the function decreases.

From Taylor's theorem

$$\begin{aligned} f(x_k + \alpha p) &= f(x_k) + \alpha p^{\mathrm{T}} \nabla f(x_k) + \frac{1}{2}\alpha^2 p^{\mathrm{T}} \nabla^2 f(x_k + tp)p, \; t \in (0, \alpha) \\ &= f(x_k) + \alpha \underbrace{p^{\mathrm{T}} \nabla f(x_k)}_{<0} + O(\alpha^2) \end{aligned}$$

Thus for $\alpha > 0$ small enough, $f(x_k + \alpha p) < f(x_k)$ implies

$$p^{\mathrm{T}} \nabla f(x_k) = \|p\| \|\nabla f(x_k)\| \cos\theta < 0 \Leftrightarrow |\theta| > \pi/2,$$

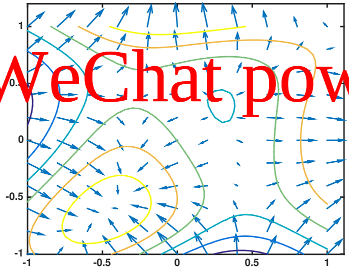
where θ is the angle between p and $\nabla f(x_k)$.

Steepest descent direction p

$$\min_p p^{\mathrm{T}} \nabla f(x_k), \qquad \text{subject to } \|p\| = 1.$$

$$\min_p p^{\mathrm{T}} \nabla f(x_k) = \min_p \|p\| \|\nabla f(x_k)\| \cos(\theta) = -\|\nabla f(x_k)\|,$$

attained for $\cos(\theta) = -1$ and $p = -\nabla f(x_k)/\|\nabla f(x_k)\|$, where θ is the angle between p and $\nabla f(x_k)$.



Newton direction

Consider the second order Taylor polynomial

$$f(x_k + p) \approx f(x_k) + p^{\mathrm{T}} \nabla f(x_k) + \frac{1}{2} p^{\mathrm{T}} \nabla^2 f(x_k) p =: m_2(p)$$

and assume $\nabla^2 f(x_k)$ is positive definite.

Newton direction minimises the second order Taylor polynomial m_2. Setting

$$m_2'(p) = \nabla^2 f(x_k) p + \nabla f(x_k) = 0,$$

yields

$$p = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k).$$

The Newton direction is reliable when $m_2(p)$ is a close approximation to $f(x_k + p)$ i.e. $\nabla^2 f(x_k + tp)$, $t \in (0, 1)$ and $\nabla^2 f(x_k)$ are close. This is the case if $\nabla^2 f$ is sufficiently smooth and the difference is of order $\mathcal{O}(\|p\|^3)$.

$$p^{\mathrm{T}} \nabla f(x_k) = -p^{\mathrm{T}} \nabla^2 f(x_k) p \leq -\sigma \|p\|^2$$

for some $\sigma > 0$. Thus unless $\nabla f(x_k) = 0$ (and hence $p = 0$), $p^{\mathrm{T}} \nabla f(x_k) < 0$ and p is a descend direction.

The step length 1 is optimal for $f(x_k + p) = m_2(p)$, thus 1 is used unless it does not produce a satisfactory reduction of f.

If $\nabla^2 f(x_k)$ is not positive definite, the Newton direction may not be defined: if $\nabla^2 f(x_k)$ is singular, $\nabla^2 f(x_k)^{-1}$ does not exist. Otherwise, p may not be a descent direction which can be remedied.

Fast local convergence (quadratic) close to the solution.

Computing the Hessian is expensive.

Quasi-Newton direction

Use symmetric positive definite (s.p.d.) approximation B_k to the Hessian $\nabla^2 f(x_k)$ in the Newton step

$$p = -B_k^{-1}\nabla f(x_k), \quad \nabla f(x_k)^{\mathrm{T}} B_k^{-1} \nabla f(x_k) > 0,$$

such that superlinear convergence is retained.

B_k is updated in each step taking into account the additional information gathered during that step. The updates make use of the fact that changes in gradient provide information about the second derivative of f along the search direction.

Secant equation

$$\nabla f(x_k + p) = \nabla f(x_k) + B_k p$$

This equation is underdetermined, different methods quasi-Newton methods differ in the way they solve it.

Given the search direction p the optimal reduction of the f amounts to minimising the function of one variable

$$\phi(\alpha) := f(x_k + \alpha p), \quad \alpha > 0.$$

This is in general to expensive (even the local minimiser), hence *inexact* line search is of interest.

Choice of step size is important. To small steps mean slow convergence, to large steps may not lead to reduction of the objective function f.

Conditions for decrease

Simple condition: require $f(x_k + \alpha p) < f(x_k)$.

Consider a sequence $f(x_k) = 5/k, k = 1, 2, \dots$. This sequence is decreasing but its limiting value is 0, while the minimum of a convex function can be smaller than 0.

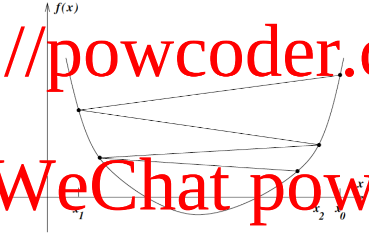


Figure: Nocedal Wright Fig 3.2

The decrease is insufficient to converge to the minimum of a convex function. Hence we need conditions for sufficient decrease.

Armijo condition

$$f(x_k + \alpha p) \leq f(x_k) + c_1 \alpha p^{\mathrm{T}} \nabla f(x_k) =: \ell(\alpha),$$

for some $c_1 \in (0,1)$. [Typically small, $c_1 = 10^{-4}$]

$\ell(\alpha)$ is a linear function with negative slope $c_1 p^{\mathrm{T}} \nabla f(x_k) < 0$,

$\ell(\alpha) = f(x_k) + c_1 \alpha p^{\mathrm{T}} \nabla f(x_k) > f(x_k) + \alpha p^{\mathrm{T}} \nabla f(x_k) = \phi(0) + \alpha\phi'(0).$

From Taylor Thm $\phi(\alpha) = \phi(0) + \alpha\phi'(0) + \alpha^2\phi''(\xi)$, $\xi \in (0, \alpha)$,
thus for sufficiently small $\alpha > 0$, $\ell(\alpha) > \phi(\alpha)$.

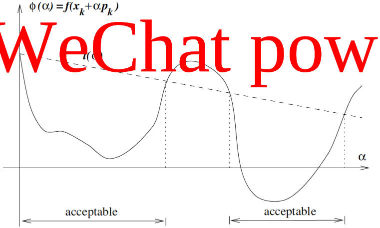


Figure: Nocedal Wright Fig 3.3

Curvature condition

Armijo condition is satisfied for all sufficiently small α, so we need another condition to avoid very small steps.

Curvature condition

$$\underbrace{p^{\mathrm{T}}\nabla f(x_k + \alpha p)}_{\phi'(\alpha)} \geq c_2 \underbrace{p^{\mathrm{T}}\nabla f(x_k)}_{\phi'(0)}, \quad c_2 \in (c_1, 1).$$

- Ensures, that we progress far enough along a *good* direction p.
- If $\phi'(\alpha)$ is strongly negative, there is a good prospect of significant decrease along p.
- If $\phi'(\alpha)$ is slightly negative (or even positive) we have a prospect of little decrease and hence we terminate the line search.
- Typically $c_2 = 0.9$ for a Newton or quasi Newton direction, $c_2 = 0.1$ for nonlinear conjugate gradient.

Curvature condition

$$\underbrace{p^{\mathrm{T}}\nabla f(x_k+\alpha p)}_{\phi'(\alpha)} \geq c_2 \underbrace{p^{\mathrm{T}}\nabla f(x_k)}_{\phi'(0)}, \quad c_2 \in (c_1, 1).$$

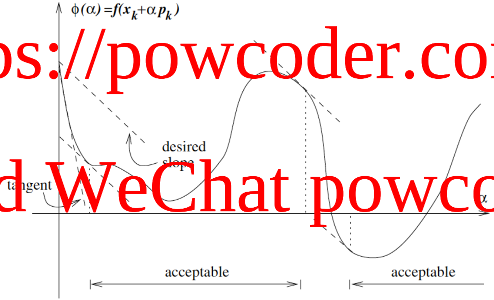


Figure: Nocedal Wright Fig 3.4

Wolfe conditions

The sufficient decrease (Armijo rule) and curvature conditions together are called **Wolfe conditions**

$$
\begin{aligned}
f(x_k + \alpha p) &\leq f(x_k) + c_1 \alpha p^{\mathrm{T}} \nabla f(x_k), \\
p^{\mathrm{T}} \nabla f(x_k + \alpha p) &\geq c_2 p^{\mathrm{T}} \nabla f(x_k),
\end{aligned}
$$

for $0 < c_1 < c_2 < 1$.

Possibly includes points far away from stationary points, hence **strong Wolfe conditions** to disallow "too positive" values of $\phi'(\alpha)$

$$
\begin{aligned}
f(x_k + \alpha p) &\leq f(x_k) + c_1 \alpha p^{\mathrm{T}} \nabla f(x_k), \\
|p^{\mathrm{T}} \nabla f(x_k + \alpha p)| &\leq c_2 |p^{\mathrm{T}} \nabla f(x_k)|,
\end{aligned}
$$

for $0 < c_1 < c_2 < 1$.

Let $f : \mathbb{R}^n \to \mathbb{R}$ continuously differentiable and p be a descent direction at x_k. If f is bounded below along the ray $\{x_k + \alpha p \mid \alpha > 0\}$, then there exists an interval of step lengths satisfying both the Wolfe conditions and strong Wolfe conditions.

Proof:

For $\alpha > 0$, $\phi(\alpha) = f(x_k + \alpha p)$ is bounded below while $\ell(\alpha) = f(x_k) + \alpha c_1 p^{\mathrm{T}} \nabla f(x_k)$ is unbounded below as $c_1 p^{\mathrm{T}} \nabla f(x_k) < 0$ and for small α, $\ell(\alpha) > \phi(\alpha)$ as $c_1 < 1$. Thus $\ell(\alpha)$ has to intersect $\phi(\alpha)$ at least once. Let α' be the smallest value for which

$$\phi(\alpha') = f(x_k + \alpha' p) = f(x_k) + \alpha' c_1 p^{\mathrm{T}} \nabla f(x_k) = \ell(\alpha').$$

Then the sufficient decrease condition holds for all $\alpha \leq \alpha'$.

Furthermore, by the mean value theorem

$$\exists \alpha'' \in (0, \alpha') : \; f(x_k + \alpha' p) - f(x_k) = \alpha' p^{\mathrm{T}} \nabla f(x_k + \alpha'' p)$$

and we obtain

$$p^{\mathrm{T}} \nabla f(x_k + \alpha'' p) = c_1 p^{\mathrm{T}} \nabla f(x_k) > c_2 p^{\mathrm{T}} \nabla f(x_k)$$

since $c_1 < c_2$ and $p^{\mathrm{T}} \nabla f(x_k) < 0$ and therewith α'' satisfies Wolfe conditions. As the inequality in curvature condition for α'' holds strictly, by continuity of ∇f, the inequality (and hence Wolfe conditions) also holds in an interval containing α''. Furthermore, as all terms in the last equation are negative strong Wolfe conditions hold for the same interval.

Wolfe conditions are scale-invariant in the sense that are unaffected by scaling the function or affine change of variables. They can be used in most line search methods and are particularly important for quasi-Newton methods.

Goldstein conditions

$$f(x_k) + (1-c)\alpha p^{\mathrm{T}} \nabla f(x_k) \le f(x_k + \alpha p) \le f(x_k) + c\alpha p^{\mathrm{T}} \nabla f(x_k)$$

with $0 < c < 1/2$.

The second inequality is the sufficient decrease condition. The first inequality controls the step length from below. Disadvantage w.r.t. Wolfe conditions is that it can exclude all minimisers of ϕ.

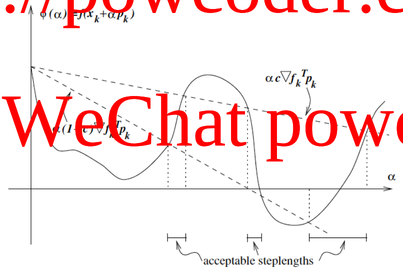


Figure: Nocedal Wright Fig 3.6

Backtracking line search

Choose $\bar{\alpha} > 0, \rho \in (0,1), c \in (0,1)$
Set $\alpha = \bar{\alpha}$
repeat
  $\alpha = \rho\alpha$
until $f(x_k + \alpha p) \leq f(x_k) + c\alpha p^{\mathrm{T}} \nabla f(x_k)$

- Terminates in finite number of steps: α will eventually become small enough to satisfy sufficient decrease condition.
- Prevents too short step lengths: the accepted α is within factor ρ of the previous value, α/ρ, which was rejected for violating the sufficient decrease condition i.e. being too long.
- ρ can vary in $[\rho_{\min}, \rho_{\max}] \subset (0,1)$ between iterations.
- In Newton and quasi-Newton methods $\bar{\alpha} = 1$, but different values can be appropriate for other algorithms.
- Well suited for Newton methods, less appropriate for quasi-Newton and conjugate gradient methods.

Consider an iteration

$$x_{k+1} = x_k + \alpha_k p_k, \;\; k = 0, 1, \ldots,$$

where p_k is a descent direction and α_k satisfied the Wolfe conditions.

Let f be bounded below in $\mathbb{R}^n$ and continuously differentiable in an open set $\mathcal{M}$ containing the level set $\{x : f(x) \leq f(x_0)\}$.

If ∇f is Lipschitz continuous on $\mathcal{M}$ i.e.

$$\exists L > 0 : \;\; \|\nabla f(x) - \nabla f(\bar{x})\| \leq L\|x - \bar{x}\|, \;\; \forall x, \bar{x} \in \mathcal{M}$$

then

$$\sum_{k \geq 0} \cos^2 \theta_k \|\nabla f(x_k)\|^2 < \infty,$$

where $\theta_k = \angle(p_k, -\nabla f(x_k))$.

Subtracting $p_k^{\mathrm{T}}\nabla f(x_k)$ from both sides of curvature condition

$$p_k^{\mathrm{T}}\nabla f(\underbrace{x_k+\alpha p_k}_{=x_{k+1}}) \ge c_2 p_k^{\mathrm{T}}\nabla f(x_k)$$

we obtain

$$p_k^{\mathrm{T}}(\nabla f(x_{k+1}) - \nabla f(x_k)) \ge (c_2 - 1) p_k^{\mathrm{T}}\nabla f(x_k).$$

On the other hand the Lipschitz condition implies

$$p_k^{\mathrm{T}}(\nabla f(x_{k+1}) - \nabla f(x_k)) \le \|\nabla f(x_{k+1}) - \nabla f(x_k)\|\|p_k\| \le \alpha_k L\|p_k\|^2.$$

Combining the two inequalities we obtain a lower bound on the step size

$$\alpha_k \ge \frac{c_2 - 1}{L}\frac{p_k^{\mathrm{T}}\nabla f(x_k)}{\|p_k\|^2}.$$

Substituting this inequality into the sufficient decrease condition

$$f(x_{k+1}) \le f(x_k) + c_1 \frac{c_2 - 1}{L} \frac{(p_k^{\mathrm{T}} \nabla f(x_k))^2}{\|p_k\|^2}$$

with $\cos\theta_k = -\frac{p_k^{\mathrm{T}} \nabla f(x_k)}{\|\nabla f(x_k)\| \|p_k\|}$ yields

$$f(x_{k+1}) \le f(x_k) - c \cos^2\theta_k \|\nabla f(x_k)\|^2,$$

where $c = c_1(1 - c_2)/L$.

Summing over all indices up to k we obtain

$$f(x_{k+1}) \le f(x_0) - c \sum_{j=0}^{k} \cos^2\theta_j \|\nabla f(x_j)\|^2$$

and since f is bounded from below

$$\sum_{j=0}^{k} \cos^2\theta_j \|\nabla f(x_j)\|^2 \le (f(x_0) - f(x_{k+1}))/c < C$$

where $C > 0$ is some positive constant. Taking limits $\sum_{k=0}^{\infty} \cos^2\theta_k \|\nabla f(x_k)\|^2 < \infty$.

Global convergence

Goldstein or strong Wolfe conditions also imply the Zoutendijk condition

$$\sum_{k \geq 0} \cos^2 \theta_k \|\nabla f(x_k)\|^2 < \infty.$$

The Zoutendijk condition implies

$$\cos^2 \theta_k \|\nabla f(x_k)\|^2 \to 0,$$

which can be used to derive *global* convergence results for line search algorithms.

If the method ensures that $\cos \theta_k \geq \delta > 0$, $\forall k$ i.e. θ_k is bounded away from $\pi/2$, it follows that

$$\lim_{k \to \infty} \|\nabla f(x_k)\| = 0.$$

This is the strongest global convergence result that can be obtained for such iteration (convergence to a stationary point) without additional assumptions.

In particular, the steepest descent ($p_k = -\nabla f(x_k)$) produces a gradient sequence which converges to 0 if it uses a line search satisfying Wolfe or Goldstein conditions.

For some algorithms e.g. nonlinear conjugate gradient methods, only a weaker result can be obtained

$$\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0$$

i.e. only subsequence of gradient norms $\|\nabla f(x_{k_i})\|$ converges to 0 rather than the whole sequence.

Those limits can be proved by contradiction:
Suppose that $\|\nabla f(x_k)\| \geq \gamma$ for some $\gamma > 0$ for all k sufficiently large. Then from $\cos^2\theta_k \|\nabla f(x_k)\|^2 \to 0$ we conclude that $\cos\theta_k \to 0$ i.e. the entire sequence $\{\cos\theta_k\}$ converges to 0.
Thus to show the weak convergence result it is enough to show that a subsequence $\{\cos\theta_{k_j}\}$ is bounded away from 0.

Consider any algorithm which
(i) decreases the objective function in each iteration,
(ii) every mth iteration is a steepest descent step with step length satisfying the Wolfe or Goldstein conditions.
Since $\cos\theta_k = 1$ for steepest descent steps, this provides the subsequence bounded away from 0. The algorithm can do something better in remaining $m-1$ iterates, while the occasional steepest descent step will guarantee the overall (weak) global convergence.

Rate of convergence

Unfortunately, rapid convergence sometimes conflicts with global convergence.

Example: Steepest descent is globally convergent (with appropriate step sizes) but it can be very slow in practice. On the other hand, while Newton iteration converges rapidly when we are close to the solution, the Newton step may not even be a descent direction far away from the solution.

The challenge: design algorithms with good global convergence properties and rapid convergence rate.

Steepest descent with exact line search for strictly convex quadratic function

$$f(x) = \frac{1}{2}x^{\mathrm{T}}Qx - b^{\mathrm{T}}x$$

where Q is symmetric positive definite.

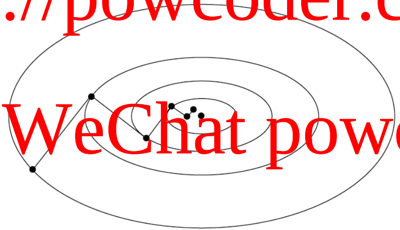


Figure: Nocedal Wright Fig 3.7

Steepest descent

Characteristic zig-zag due to elongated shape of the ellipse. If the level sets were circles instead, the steepest descent would need one step only.

Convergence rate of steepest descent with exact line search:

$$\underbrace{\|x_{k+1} - x^\star\|_Q^2}_{=f(x_{k+1})-f(x^\star)} \le \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 \underbrace{\|x_k - x^\star\|_Q^2}_{=f(x_{k+1})-f(x^\star)},$$

where $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ are the eigenvalues of Q, and $\|x\|_Q^2 = x^{\mathrm{T}} Q x$. Note: for quadratic strictly convex function we obtain objective function and (for free) iterate convergence rates!

The objective function convergence rate is essentially the same for steepest descent with exact line search when applied to a twice continuously differentiable nonlinear function satisfying sufficient conditions at $x^\star$.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable with Lipschitz continuous Hessian in a neighbourhood of the solution $x^\star$ satisfying the sufficient conditions. Note that the Hessian $\nabla^2 f$ is positive definite also in the vicinity of the solution $x^\star$.

The iterates x_k computed by the Newton method (note step length 1)

$$x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$$

converge locally quadratically i.e. for starting point x_0 sufficiently close to $x^\star$.

The sequence of gradient norms $\|\nabla f(x_k)\|$ also converges quadratically to 0.

Local convergence: note that away from the solution ∇f_k may not be positive definite and hence p_k may not be a descent direction. Global convergence with Hessian modification is discussed later.

Convergence rates: Newton-type methods

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable.
Let $\{x_k\}$ be a sequence generated by a descent method

$$x_{k+1} = x_k + \alpha p_k$$

for step sizes satisfying Wolfe conditions with $c_1 \leq 1/2$.
If the sequence $\{x_k\}$ converges to a point $x^\star$ satisfying the sufficient conditions and the search direction satisfies

$$\lim_{k\to\infty} \frac{\|\nabla f(x_k) + \nabla^2 f(x_k) p_k\|}{\|p_k\|} = 0,$$

then for all $k > k_0$, the step length $\alpha_k = 1$ is admissible, and for that choice of $\alpha_k = 1$, $k > k_0$, the sequence $\{x_k\}$ converges to $x^\star$ superlinearly.

Note: once close enough to the solution so that $\nabla^2 f(x_k)$ became s.p.d., the limit is trivially satisfied and for $\alpha_k = 1$ we recover local quadratic convergence.

Convergence rates: quasi-Newton methods

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable.
Let $\{x_k\}$ be a sequence generated by a quasi-Newton method (note step length 1, B_k s.p.d.)

$$x_{k+1} = x_k \underbrace{- B_k^{-1} \nabla f(x_k)}_{p_k}.$$

Assume the sequence $\{x_k\}$ converges to a point $x^\star$ satisfying the sufficient conditions. Then $\{x_k\}$ converges superlinearly if an only if

$$\lim_{k \to \infty} \frac{\|(B_k - \nabla^2 f(x^\star))p_k\|}{\|p_k\|} = 0.$$

Note: the superlinear convergence rate can be attained even if the sequence $\{B_k\}$ does not converge to $\nabla^2 f(x^\star)$. It suffices that B_k becomes increasingly accurate approximation to $\nabla^2 f(x^\star)$ along the search direction p_k. Quasi-Newton methods use it to construct B_k.

Hessian modifications

Away from the solution, the Hessian may not be positive definite, and the Newton direction may not be a descent direction. The general solution is to consider positive definite approximations.

$B_k = \nabla^2 f(x_k) + E_k$, where E_k is chosen to ensure that B_k is sufficiently positive definite.

Global convergence results can be established for Newton method with Hessian modification and step satisfying Wolfe or Goldstein or Armijo backtracking conditions provided that:

$\kappa(B_k) = \|B_k\|\|B_k^{-1}\| \leq C$ for some $C > 0$ and all k whenever the sequence of the Hessians $\{\nabla^2 f(x_k)\}$ is bounded.

Eigenvalue decomposition

$$\nabla^2 f(x_k) = Q\Lambda Q^{\mathrm{T}} = \sum_{i=1}^{n} \lambda_i q_i q_i^{\mathrm{T}}.$$

Example:

$$\nabla^2 f(x_k) = \mathrm{diag}(10, 3, -1), \quad \nabla f(x_k) = (1, -3, 2)^{\mathrm{T}}$$

$$Q = I, \quad \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3)$$

The Newton step: $p_k = (-0.1, 1, 2)^{\mathrm{T}}$

As $p_k^{\mathrm{T}} \nabla f(x_k) > 0$, it is not a descent direction.

Eigenvalue modifications (not practical):
Replace all negative eigenvalues with $\delta = \sqrt{\mathbf{u}} = 10^{-8}$, where $\mathbf{u} = 10^{-16}$ is the machine precision.

$$B_k = \sum_{i=1}^{2} \lambda_i q_i q_i^{\mathrm{T}} + \delta q_3 q_3^{\mathrm{T}} = \mathrm{diag}(10, 3, 10^{-8})$$

B_k is s.p.d. and curvature along q_1, q_2 is preserved, however the direction is dominated by q_3:

$$p_k = -B_k^{-1}\nabla f_k = -\sum_{i=1}^{2} \frac{1}{\lambda_i} q_i q_i^{\mathrm{T}} \nabla f_k - \frac{1}{\delta} q_3 q_3^{\mathrm{T}} \nabla f_k \approx -(2 \times 10^8) q_3.$$

p_k is a descent direction but the length is very large, not in line with local validity of the Newton approximation. Thus p_k may be ineffective.
Adapt choice of δ to avoid excessive lengths. Even $\delta = 0$ which eliminates direction q_3.

Let A is symmetric $A = Q\Lambda Q^{\mathrm{T}}$.
The correction matrix ΔA of minimum Frobenius norm that ensures $\lambda_{\min}(A + \Delta A) \geq \delta$ is given by

$$\Delta A = Q \operatorname{diag}(\tau_i) Q^{\mathrm{T}}, \quad \text{with} \quad \tau_i = \begin{cases} 0, & \lambda_i \geq \delta \\ \delta - \lambda_i, & \lambda_i < \delta. \end{cases}$$

and the modified matrix is

$$A + \Delta A = Q(\Lambda + \operatorname{diag}(\tau_i)) Q^{\mathrm{T}}.$$

Frobenius norm is defined $\|A\|_F^2 = \sum_{i,j=1}^n a_{ij}^2 = \sum_{i=1}^n \lambda_i^2$.

The correction matrix ΔA of minimum Euclidean norm that satisfies $\lambda_{\min}(A + \Delta A) \geq \delta$ is given by

$$\Delta A = \tau I, \quad \text{with} \quad \tau = \max(0, \delta - \lambda_{\min}(A)).$$

and the modified matrix has the form $A + \tau I$.

Cholesky factorisation of $A + \tau I$

Simple idea

If $\min_i a_{ii} \leq 0$ set $\tau_0 = -\min_i a_{ii} + \beta$ for some small $\beta > 0$ (e.g. 10^{-3}), otherwise $\tau_0 = 0$.

Attempt the Cholesky algorithm to obtain $LL^{\mathrm{T}} = A + \tau_k I$

If not successful, increase $\tau_{k+1} = \max(2\tau_k, \beta)$ and reattempt.

Drawback: possibly multiple failed attempts to factorise.

Consider the case $n = 3$. The equation $A = LDL^T$ is given by

$$\begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix} \begin{bmatrix} 1 & l_{21} & l_{31} \\ 0 & 1 & l_{32} \\ 0 & 0 & 1 \end{bmatrix}.$$

(The notation indicates that A is symmetric.) By equating the elements of the first column, we have

$$\begin{aligned} a_{11} &= d_1, \\ a_{21} &= d_1 l_{21} &\Rightarrow\quad l_{21} &= a_{21}/d_1, \\ a_{31} &= d_1 l_{31} &\Rightarrow\quad l_{31} &= a_{31}/d_1. \end{aligned}$$

Proceeding with the next two columns, we obtain

$$\begin{aligned} a_{22} &= d_1 l_{21}^2 + d_2 &\Rightarrow\quad d_2 &= a_{22} - d_1 l_{21}^2, \\ a_{32} &= d_1 l_{31} l_{21} + d_2 l_{32} &\Rightarrow\quad l_{32} &= \left(a_{32} - d_1 l_{31} l_{21}\right)/d_2, \\ a_{33} &= d_1 l_{31}^2 + d_2 l_{32}^2 + d_3 &\Rightarrow\quad d_3 &= a_{33} - d_1 l_{31}^2 - d_2 l_{32}^2. \end{aligned}$$

Figure: Nocedal Wright Ex. 3.1

Cholesky decomposition of indefinite matrix

For A indefinite:

- The factorisation $A = LDL^{\mathrm{T}}$ may not exist.
- Even if it does exist, the algorithm can be unstable i.e. elements of L and D can become arbitrarily large.
- Posterior modification of D to force the elements to be positive may break down or result in a matrix very different to A.
- Instead, modify A during the factorisation to achieve that the elements of D are sufficiently positive and the elements of L and D are not to large.

Modified Cholesky decomposition

Choose $\delta, \beta > 0$. While computing jth column of L, D ensure

$$d_j \geq \delta, \quad |m_{ij}| \leq \beta, \quad i = j+1, j+2, \dots, n,$$

where $m_{ij} = l_{ij}\sqrt{d_j}$.

To satisfy these bounds we only need to change how d_j is computed, from $d_j = c_{jj}$ to

$$d_j = \max\left(|c_{jj}|, \frac{\theta_j^2}{\beta^2}, \delta\right), \quad \text{with } \theta_j = \max_{j<i\leq n} |c_{ij}|,$$

where $c_{ij} = l_{ij} d_j$. Note: θ_j can be computed before d_j because computing $c_{ij}, j < i \leq n$ only needs previous columns!

Verification:
$d_j \geq \delta$ due to taking maximum

$$|m_{ij}| = |l_{ij}\sqrt{d_j}| = \frac{|c_{ij}|}{\sqrt{d_j}} \leq \frac{|c_{ij}|\beta}{\theta_j} \leq \beta, \quad \forall i > j.$$

Properties:

- Modifies the Hessian during factorization where necessary.
- The modified Cholesky factors exist and are bounded relative to the norm of the actual Hessian.
- It does not modify Hessian if it is sufficiently positive definite.

This is the basis for the modified Cholesky factorisation which also introduces symmetric row and column permutations to reduce the size of the modification.

$$PAP^{\mathrm{T}} + E = LDL^{\mathrm{T}} = MM^{\mathrm{T}},$$

where E is a nonnegative diagonal matrix that is zero if A is sufficiently positive definite.

It has been shown, that the matrices obtained by this modified Cholesky algorithm to the exact Hessian $\nabla^2 f(x_k)$ have bounded condition numbers, hence some global convergence results can be obtained.

How to find a step length satisfying one of the termination conditions e.g. Wolfe etc. for

$$\phi(\alpha) = f(x_k + \alpha p_k),$$

where p_k is a descent direction i.e. $\phi'(0) = p_k^{\mathrm{T}} f(x_k) < 0$.

If f is a convex quadratic function $f(x) = \frac{1}{2}x^{\mathrm{T}}Qx - b^{\mathrm{T}}x$, it has a global minimiser along the ray $x_k + \alpha p_k$ which can be calculated analytically

$$\alpha_k = -\frac{p_k^{\mathrm{T}} \nabla f(x_k)}{p_k^{\mathrm{T}} Q p_k}.$$

For general nonlinear functions iterative approach is necessary.

Line search algorithms can be classified according to the information they use:

Methods using only function evaluations can be very inefficient as they need to continue iterating until a very small interval has been found.

Methods using gradient information can determine wether the current step length satisfies e.g. Wolfe or Goldstein conditions which require gradients to evaluate.

Typically, they consist of two phases: *bracketing phase* which finds an interval containing acceptable step lengths and the *selection phase* which locates the final step in the interval.

Aim: find a step length α that satisfies sufficient decrease condition without being to small. [similarity to backtracking]

Sufficient decrease condition

$$\phi(\alpha_k) \leq \phi(0) + c_1 \alpha_k \phi'(0)$$

.

We want to compute as few derivatives $\nabla f(x)$ as possible.

Initial guess α_0: check sufficient decrease condition

$$\phi(\alpha_0) \leq \phi(0) + c_1 \alpha_0 \phi'(0).$$

If satisfied terminate the search. Otherwise, $[0, \alpha_0]$ contains acceptable step lengths.

Quadratic approximation $\phi_q(\alpha)$ to ϕ by interpolating the available information: $\phi_q(0) = \phi(0)$, $\phi_q'(0) = \phi'(0)$ and $\phi_q(\alpha_0) = \phi(\alpha_0)$ yields

$$\phi_q(\alpha) = \frac{\phi(\alpha_0) - \phi(0) - \alpha_0\phi'(0)}{\alpha_0^2}\alpha^2 + \phi'(0)\alpha + \phi(0).$$

The new trial value α_1 is defined as the minimiser of ϕ_q i.e.

$$\alpha_1 = -\frac{\phi'(0)\alpha_0^2}{2(\phi(\alpha_0) - \phi(0) - \phi'(0)\alpha_0)}$$

If sufficient decrease condition is satisfied, terminate search. Otherwise, construct cubic interpolating the four available values: $\phi_c(0) = \phi(0)$, $\phi_c'(0) = \phi'(0)$, $\phi_c(\alpha_0) = \phi(\alpha_0)$ and $\phi_c(\alpha_1) = \phi(\alpha_1)$ yields

$$\phi_c(\alpha) = a\alpha^3 + b\alpha^2 + \alpha\phi'(0) + \phi(0),$$

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{\alpha_0^2\alpha_1^2(\alpha_1 - \alpha_0)} \begin{bmatrix} \alpha_0^2 & -\alpha_1^2 \\ -\alpha_0^3 & \alpha_1^3 \end{bmatrix} \begin{bmatrix} \phi(\alpha_1) - \phi(0) - \phi'(0)\alpha_1 \\ \phi(\alpha_0) - \phi(0) - \phi'(0)\alpha_0 \end{bmatrix}$$

By differentiating ϕ_c we find the minimiser $\alpha_2 \in [0, \alpha_1]$

$$\alpha_2 = \frac{-b + \sqrt{b^2 - 3a\phi'(0)}}{3a}.$$

If necessary repeat the cubic interpolation with $\phi_c(0) = \phi(0)$, $\phi_c'(0) = \phi'(0)$ and the two most recent values $\phi_c(\alpha_{k-1}) = \phi(\alpha_{k-1})$ and $\phi_c(\alpha_k) = \phi(\alpha_k)$ until α_{k+1} satisfies the sufficient decrease condition.

Safeguard: If any α_i is either too close to α_{i-1} or much smaller than α_{i-1} we reset $\alpha_i = \alpha_{i-1}/2$.

If derivatives can be computed along the function values at little additional cost, we can also devise variant interpolating ϕ, ϕ' at two most recent values.

Initial step length

For Newton and quasi Newton $\alpha_0 = 1$. This ensures that unit step length will be taken whenever they satisfy the termination conditions and allow for quick convergence.

For methods which do not produce well scaled search direction like steepest descent or conjugate gradient it is important to use available information to make the initial guess e.g.:

- First order change in function at iterate x_k will be the same as that obtained at previous step
  i.e. $\alpha_0 p_k^{\mathrm{T}} \nabla f(x_k) = \alpha_{k-1} p_{k-1}^{\mathrm{T}} \nabla f(x_{k-1})$

$$\alpha_0 = \alpha_{k-1} \frac{p_{k-1}^{\mathrm{T}} \nabla f(x_{k-1})}{p_k^{\mathrm{T}} \nabla f(x_k)}.$$

- Interpolate quadratic to data $f(x_{k-1}), f(x_k)$ and $p_{k-1}^{\mathrm{T}} \nabla f(x_{k-1})$ and define α_0 to be its minimiser

$$\alpha_0 = \frac{2(f(x_k) - f(x_{k-1}))}{\phi'(0)}.$$

It can be shown that if $x_k \to x^\star$ superlinearly, then the ratio converges to 1. If we adjust by setting $\alpha_0 = \min(1, 1.01\alpha_0)$ we find that the unit step length will eventually always be tried and accepted and the superlinear convergence will be observed.